JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 4
No. 2
Desember
2018

Perbandingan Kinerja Hasil Seleksi Fitur pada Prediksi Kinerja Akademik Siswa Berbasis Pohon Keputusan

Achmad Shoddiq Bayu Asmoro[#1], Wahyu Sakti Gunawan Irianto[#2], Utomo Pujianto[#3]

[#]*Program Studi Informatika Universitas Negeri Malang*
Jalan Semarang No. 5, Malang, Jawa Timur

[1]bayuasmoro552@gmail.com
[2]iriantowsg@yahoo.com
[3]utomo.pujianto.ft@um.ac.id

***Abstrak*—Sistem manajemen *E-learning* merupakan bentuk kemajuan teknologi dalam bidang pendidikan dan telah banyak menghasilkan kumpulan data-data pendidikan yang salah satunya adalah data aktivitas pembelajaran siswa dalam sistem manajemen *E-learning*. Banyaknya data pendidikan yang belum tereksplorasi dengan baik dapat di manfaatkan dengan menggunakan teknik data mining. Pada penelitian ini akan dilakukan perbandingan 3 model data berbeda yaitu data awal tanpa *preprocessing* dan data yang di *preprocessing* menggunakan seleksi fitur *correlation-based feature selection* dan *Information Gain*. Data yang digunakan adalah data aktivitas pembelajaran siswa dalam sistem manajemen *E-learning*. Selanjutnya proses pengujian data dengan menggunakan 10 *folds cross validation* dengan metode C4.5 dan evaluasi data menggunakan *confusion matrix*. Hasil dari pengujian data menggunakan algoritma C4.5 yang dikombinasikan dengan seleksi fitur *correlation-based feature selection* menghasilkan nilai akurasi yang lebih tinggi dengan nilai akurasi sebesar 76.92%. Sementara itu hasil dari pengujian data awal tanpa seleksi fitur dan data yang di seleksi fitur menggunakan *information gain* memiliki nilai akrasi yang sama dengan nilai akurasi sebesar 76.19%. Hal ini dikarenakan data yang diproses menggunakan algoritma C4.5 tanpa *preprocessing* dan data yang telah di *preprocessing* menggunakan *information gain* sama-sama menghitung nilai gain untuk membuat model pohon keputusan, dan menghasilkan model pohon keputusan yang sama. Sehingga hasil dari proses pengujian data memiliki nilai akurasi yang sama.**

Kata kunci*—Educational data mining, C4.5, Coorelation-based feature selection, Information gain.

I. PENDAHULUAN

Berkembangnya teknologi informasi dalam bidang pengolahan data banyak membawa pengaruh positif bagi dunia pendidikan. Salah satunya adalah dengan penggunaan sistem manajamen *e-learning* untuk membantu proses belajar mengajar dan proses evaluasi kinerja akademik siswa [1]. Akhir-akhir ini mulai banyak penelitian yang mengembangkan penggunaan teknik data mining dalam bidang pendidikan.

Data mining dalam bidang pendidikan berfokus pada pengembangan eksplorasi tipe data yang unik, yang salah satu pemanfaatanya dapat dikembangkan dengan menggunakan metode dari algoritma data mining, *machine-learning* dan statistika. Salah satu metode dalam bidang data mining yang sering digunakan untuk memprediksi kinerja akademik siswa adalah metode klasifikasi [2].

Metode klasifikasi merupakan suatu metode yang digunakan untuk pengelompokan data kedalam kelas yang telah ditentukan [3]. Pada penelitian ini akan digunakan suatu fitur baru yaitu: Fitur perilaku siswa. Fitur ini berhubungan dengan aktivitas interaksi siswa dengan sistem manajemen *e-learning*.

Langkah pertama yang akan dilakukan adalah *preprocessing* data. Hal ini ditujukan untuk perbaikan data agar pembelajaran algoritma menjadi lebih efektif dalam melakukan komputasi data. Selain itu *preprocessing* data berguna untuk menghilangkan atribut data yang tidak relevan dan berlebihan, sehingga proses belajar algoritma terhadap data menjadi semakin meningkat [4].

Seleksi fitur merupakan salah satu teknik penting untuk dilakukan dalam *preprocessing* data. Proses seleksi fitur bertujuan untuk menentukan jumlah fitur yang akan digunakan dalam menentukan kelas target serta mengurangi fitur yang tidak relevan. Berikut beberapa jenis teknik seleksi fitur yang dapat digunakan dalam *preprocessing* data yaitu: *Information gain*, *Gain ratio*, *Chi-square*, *Symmetrical uncertainty*, *Relief*, dan *Correlation-based feature selection*. Dari beberapa teknik seleksi fitur tersebut. Teknik seleksi fitur *correlation-based feature selection* merupakan algoritma seleksi fitur yang paling stabil dari semua pengujian skala perankingan tingkat densitas data, sedangkan algoritma seleksi fitur yang berbasis entropy mempunyai kecenderungan dalam memilih atribut yang sama dengan jumlah yang sama [5].

Untuk pengujian data digunakan metode data mining algoritma C4.5. Algoritma C4.5 merupakan salah satu metode yang paling sering digunakan dalam mengolah data pendidikan dan dari beberapa penelitian yang pernah dilakukan algoritma C4.5 dapat menghasilkan nilai akurasi

Submitted 23-10-2018; Revised 29-11-2018; Accepted 20-12-2018

yang lebih baik dibandingkan dengan beberapa algoritma lain. Berikut beberapa penelitian yang pernah dilakukan dan berkaitan dengan prediksi kinerja akademik siswa yaitu:

Implementasi Teknik Seleksi Fitur *Information Gain* pada Algoritma Klasifikasi *Machine Learning* untuk Prediksi Performa Akademik Siswa. Menjelaskan bahwa data yang digunakan dalam penelitian tersebut merupakan data akademik siswa yang berjumlah 395 dari wilayah Alentejo, Portugal. Dari hasil eksperimen yang telah dilakukan menunjukan bahwa penggunaan seleksi fitur *information gain* dapat mempengaruhi nilai akurasi algoritma dalam memprediksi performa akademik siswa. Hal ini ditunjukan pada pengujian data menggunakan algoritma C4.5 yang menghasilkan nilai akurasi sebesar 90.48% dari nilai akurasi sebelumnya yaitu sebesar 88.58% dari data yang belum dilakukan proses seleksi fitur. Pada algoritma ANN menunjukkan peningkatan nilai akurasi dari 88.15% menjadi 88.96%. Pada algoritma *Naïve bayes* juga menunjukan peningkatan nilai akurasi dari 85.67% menjadi 86.68%. Sementara itu dari hasil pengujian data menggunakan algoritma *Random Forest* dan SVM menunjukan penurunan nilai akurasi dari 90.43% menjadi 90.05% setelah dilakukan seleksi fitur pada algoritma *Random Forest*, dan 89.14% menjadi 88.1% pada algoritma SMV [6].

Selanjutnya Perbandingan Performansi Algoritma C4.5 dan CART dalam Klasifikasi Data Nilai Mahasiswa Prodi Teknik Komputer Politeknik Negeri Padang. Menjelaskan bahwa data yang digunakan merupakan data akademik mahasiswa pada semester pertama perkuliahan yang meliputi mata kuliah kalkulus, fisika, alpro, pengantar TI, p.DasPro, Das.El, dan matdis. Dari proses pengujian data yang telah dilakukan menghasilkan nilai akurasi sebesar 84.95% pada algoritma CART dan 85.61% pada algoritma C4.5. Hal ini dikarenakan data nilai mahasiswa merupakan data kelompok yang cocok dengan model klasifikasi yang ada pada algoritma C4.5 sedangkan algoritma CART lebih cocok apabila data yang digunakan merupakan data numerik [7].

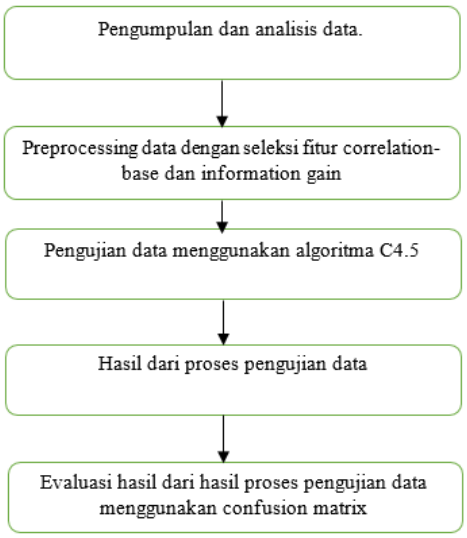
Selanjutnya penelitian dengan judul *Mining Educational Data to Predict Student's Academic Performance Using Ensemble Methods* menjelaskan perbandingan prediksi kinerja akademik siswa dengan menggunakan data fitur perilaku siswa. Data yang digunakan merupakan data interaksi siswa dengan sistem manajemen *e-learning* yang dicatat menggunakan xAPI. Selanjutnya digunakan metode *Ensemble* untuk meningkatkan kinerja klasifikasi data. Hasil dari pengujian data yang telah dilakukan menghasilkan nilai akurasi sebesar 75.8% menggunakan algoritma C4.5, 79.1% menggunakan algoritma ANN, dan 67.7% menggunakan algoritma *Naïve Bayes*. Setelah dilakukan validasi pengujian data dengan menambahkan 25 data siswa baru menghasilkan nilai akurasi sebesar 82.2% pada algoritma C4.5, 80.0% pada algoritma ANN, dan 80.0% pada algoritma Naïve Bayes [8].

Berdasarkan hasil dari beberapa penelitian yang telah dilakukan menunjukan bahwa algoritma C4.5 dapat menghasilkan nilai akurasi yang lebih baik dibandingan dengan beberapa algoritma lain seperti: ANN, SVM, *Random Forest*, dan *Naïve Bayes*. Untuk *preprocessing* data akan digunakan seleksi fitur *information gain*, pemilihan seleksi fitur *information gain* dikarenakan penggunaan seleksi fitur ini terbukti dapat menungkatkan performa algoritma dalam melakukan pengujian data[6]. Selain itu akan digunakan juga metode seleksi fitur *Correlation-base Feature Selection*. Hal ini dikarenakan *Correlation-base Feature Selection* merupakan algoritma seleksi fitur yang paling stabil dalam semua pengujian skala perankingan tingkat densitas data [5]. Atas dasar itulah dalam penelitian ini akan digunakan algoritma C4.5 yang dikolaborasikan dengan seleksi fitur *information gain* yang akan dibandingkan dengan metode C4.5 yang dikolaborasikan dengan seleksi fitur *correlation-based feature selection*.

II. METODOLOGI PENELITIAN

A. *Kerangka Penelitian*

Pada penelitian ini digunakan kerangka penelitian yang terdiri dari berbagai tahapan. Tahap pertama adalah analisis dan pengumpulan data, dimana data yang digunakan merupakan data aktivitas pembelajaran siswa dalam sistem manajemen *E-learning*. Pada tahap kedua dilakukan *preprocessing* data dimana data akan diseleksi fitur hal ini bertujuan untuk mengurangi jumlah fitur yang tdak relevan dan data yang berlebihan. Pada tahap ketiga dilakukan pengujian data menggunakan algoritma C4.5. Selanjutnya pada tahap keempat dilakukan evaluasi hasil dari pengujian data menggunakan *confusion matrix*.



Gambar 1. Kerangka Penelitian

B. *Sumber Data*

Dataset yang digunakan di dalam penelitian ini adalah dataset sekunder yang diperoleh dari website www.kaggle.com/aljarah/xAPI-Edu-Data data berupa data pendidikan yang terkait dengan aktivitas siswa dengan sistem manajemen *e-learning* [9]. Pada dataset tersebut terdapat 480 data siswa yang dicatat dalam berbagai kategori, seperti kategori fitur demografi seperti gender dan kewarganegaraan, fitur latar belakang akademis seperti tahap pendidikan, tingkat kelas, dan seksi, dan terakhir adalah fitur perilaku siswa seperti *discussion, raisehand, visited resource, announcements view dan parent school statification*. Dari 480 data yang ada, data yang digunakan sebanyak 281 data yaitu data dari siswa *middle school* dan *high school*.

C. Pengumpulan Data

Pada penelitian sebelumnya data didaptkan dari sistem LMS (*Learning Manajemen System*) yang disebut Kalboarad 360. Kalboard 360 adalah LMS multi-agen, yang dibuat untuk memfasilitas pembelajaran online yang menggunakan kemajuan teknologi terdepan dalam penerapanya. Adanya sistem pembelajaran online diharapkan dapat memudahkan pengguna dalam mengakses sumber belajar melalui perangkat apapun dengan menggunakan koneksi internet.

Data dikumpulkan dengan menggunakan alat pelacak aktivitas pelajar, yang disebut experience API (xAPI). XAPI adalah komponen dari *Training and Learning Architecture* (TLA) yang digunakan untuk memantau aktivitas pembelajaran yang dilakukan oleh peserta didik dalam LMS seperti membaca artikel atau menonton video pelatihan. XAPI dapat membantu penyedia kegiatan pembelajaran untuk memantau pelajar, aktivitas dan objek yang menggambarkan pengalaman belajar. Tujuan dari X-API digunakan untuk memantau perilaku siswa melalui proses pendidikan untuk mengevaluasi fitur-fitur yang mungkin berdampak pada prestasi akademik siswa [8].

Data yang telah didapatkan merupakan data dari siswa sekolah dasar, siswa sekolah menengah, dan siswa sekolah tingkat atas. Selanjutnya dilakukan analisis data, dimana analisis dilakukan berdasarkan teori-teori pendidikan yang menjelaskan perilaku dan pola pikir dari siswa sekolah dasar, sekolah menengah, dan sekolah atas.

Persepsi dan proses cara memecahkan suatu masalah berbeda-beda dan terus berkembang seiring dengan perkembangan anak pada usia sekolah [10]. Setiap siswa pada kelompok kelas pendidikan memiliki kemampuan dan karakter yang berbeda-beda, sehingga perlu adanya perlakukan yang berbeda pada setiap kelompok kelas pendidikan. Karakter dan pola pikir dari siswa sekolah dasar berbeda dengan siswa dari sekolah menengah dan siswa sekolah atas. Pada masa anak-anak kecenderung untuk menirukan perilaku dari orang yang diidolakan sangat besar. Sementara pada anak yang berusia remaja lebih berkeinginan untuk diakui menjadi orang dewasa dan dapat menentukan jalan hidupnya sendiri[11]. Atas dasar itulah dalam penelitian ini akan digunakan data dari siswa sekolah menengah dan siswa sekolah atas.

D. Seleksi Fitur

Pada proses seleksi fitur dilakukan analisis fitur terlebih dahulu dari data yang telah didapatkan. Selanjutnya dilakukan analisis dan seleksi fitur menggunakan *correlation-based feature selection*, dan *information gain*. Mengapa perlu dilakukan proses seleksi fitur, karena setelah melakukan studi literatur dan berdasarkan kajian teori yang telah didapatkan, proses seleksi fitur merupakan salah satu teknik *preprocessing* yang sangat penting untuk dilakukan. Hal ini ditujukan untuk mengurangi jumlah atribut yang terkait dalam menentukan kelas target. Dikarenakan atribut data yang berlebihan dan tidak relevan dapat berpengaruh pada proses pembelajaran data dan pada saat proses komputasi, sehingga mempengaruhi hasil nilai akurasi dari proses klasifikasi tersebut [4].

Dari 281 data dan 16 fitur yang ada dilakukan uji korelasi menggunakan *correlation-based* dimana uji korelasi dilakukan dengan melakukan perhitungan dan perbandingan tingkat korelasi antara suatu fitur dengan kelasnya dan fitur ke fitur lainya. *Correlation-based feature selection* merupakan suatu teknik seleksi fitur yang akan memperhitungkan suatu fitur yang memiliki korelasi yang baik dan relevan dengan kelasnya dan tidak memiliki korelasi berlebihan dengan fitur lainya [12]. Selain itu *correlation-based* akan memberikan nilai tinggi terhadap suatu fitur yang memiliki korelasi yang tinggi dengan kelasnya dan memiliki korelasi rendah dengan fitur lainya [13].

Selain menggunakan *correlation-based* data awal yang didapatkan sebanyak 281 data dan 16 fitur akan di *preprocessing* menggunakan *information gain*, *Information gain* merupakan teknik *preprocessing* data dengan melakukan pembobotan pada sebuah fitur dengan menggunakan perhitungan maksimal entropy [14]. Nilai entropy yang didapatkan digunakan untuk menentukan atribut terbaik. *Information gain* akan meranking nilai gain dari setiap atribut, dimana atribut dengan nilai gain tertinggi merupakan atribut yang paling berpengaruh dengan kelasnya dan akan dijadikan node pertama pada model *tree* yang akan dibuat [15].

E. Pengujian Data

Dalam proses pengujian data, data yang digunakan merupakan data awal yang didapatkan dan data dari hasil proses seleksi fitur menggunakan *Correlation-based Feature Selection* dan *information gain*. Data yang berjumlah 281 data akan diproses menggunakan 10-*folds cross validation* dengan metode C4.5. 10-*folds cross validation* akan membagi kumpulan data menjadi data latih dan data uji. Teknik ini akan membagi kumpulan data menjadi 10 subset dengan ukuran yang sama, sembilan dari 10 subset data digunakan untuk pelatihan, sementara satu subset yang tertinggal digunakan untuk pengujian. Proses diulang selama sepuluh kali, dan hasil akhirnya diperkirakan sebagai tingkat kesalahan rata-rata pada contoh uji [16].

F. Evaluasi Hasil Pengujian Data

Pada proses evaluasi dari hasil pengujian data, digunakan tiga ukuran evaluasi yang berbeda yaitu : Akurasi, Presisi, dan *Recall*. Tabel I menunjukan *confusion matrix* dari pengukuran.

TABEL I
CONFUSION MATRIX

		Dectected	
		Positive	***Negative***
Actual	***Positive***	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Akurasi adalah tingkat kedekatan antara nilai prediksi dengan nilai aktual. Presisi adalah tingkat ketepatan antara informasi yang diminta dengan jawaban yang diberikan oleh sistem. *Recall* adalah tingkat keberhasilan sistem dalam menemumakan kembali informasi. Berikut perhitungna Akurasi, *Precision*, dan *Recall*.

Berikut adalah rumus untuk perhitungan Akurasi, Presisi, dan *Recall*.

$$Akurasi = \frac{TP + TN}{TP+FN+FP+TN} \quad (1)$$

$$Presisi = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

III. HASIL DAN PEMBAHASAN

A. Hasil Seleksi Fitur Data Pendidikan

Proses seleksi fitur dilakukan menggunakan *correlation-based feature selection* dan *Information Gain*. Data yang didapatkan dari penelitian sebelumnya yang berjumlah 480 data dan 16 fitur yang dibagi kedalam tiga kategori yaitu: fitur demografi, fitur akademik, dan fitur perilaku siswa [8]. Dari 480 data yang ada, data yang digunakan sebanyak 281 data yaitu data dari siswa *middle school* dan *high school*.

Dari data pendidikan yang telah didapatkan selanjutnya dilakukan pemrosesan ulang untuk menyeleksi fitur data menggunakan *correlation-based feature selection*, proses seleksi fitur *correlaion-based* merupakan proses seleksi fitur dengan melakukan perhitungan dan perbandigan tingkat korelasi antara atribut dengan kelas dan atribut dengan atribut lainya. Atribut yang dipilih merupakan atribut memiliki korelasi tinggi dengan kelasnya dan memiliki tingkat korelasi rendah dengan atribut lainya. Proses seleksi fitur ini dilakukan menggunakan library dari weka yang dimasukan kedalam program java yang menghasilkan data seperti pada Tabel II.

TABEL II
CONTOH DATA SETELAH DILAKUKAN PREROCESSING MENGGUNAKAN CORRELATION-BASED

Ge	***Rel***	***Rais***	***Visi***	***Ann***	***Dis***	***Sad***	***Class***
M	Father	50	10	15	17	Under 7	M
M	Father	40	50	12	22	Above 7	M
F	Mum	60	70	44	50	Above 7	H
F	Father	60	60	33	70	Above 7	M

Selanjutnya *preprocessing* data pendidikan menggunakan *Information Gain*. *Information Gain* merupakan suatu teknik perankingan nilai suatu atribut berdasarkan perhitungan maksimal entropy, dimana atribut yang memiliki nilai gain tertinggi akan ditaruh pada baris pertama dan merupakan atribut yang paling berpengaruh terhadap kelas data. *Preprocessing* data dilakukan dengan menggunakan library dari weka yang dimasukan kedalam program java sehingga menghasilkan data perankingan atibut seperti pada Tabel III.

TABEL III
CONTOH ATRIBUT DATA SETELAH DILAKUKAN PREPROCESSING MENGGUNAKAN INFORMATION GAIN

Ranker	**Attribute**		**Ranker**	**Attribute**
0.52	Visited Resource		0.12	Parent Ans
0.46	Raisehands		0.11	Topic
0.35	Student Absence		0.09	Parent school statisfaction
0.28	Announcement View		0.047	Gender
0.20	Discussion		0.040	Grade ID
0.14	Relation		0.009	Semester
0.14	Place of birth		0.004	Section ID
0.14	Nationaly		0.0008	Stage ID

B. Hasil Proses Pengujian dan Evaluasi Data

Setelah dilakukan proses seleksi fitur hasil seleksi disimpan dengan format CSV. Selanjutnya data diambil dan dimasukan ke dalam program java yang telah di tambahkan dengan library weka untuk dilakukan proses pengujian data menggunakan 10 *folds cross validation*. Terdapat tiga model data yang akan digunakan dalam proses *cross validation* yaitu data awal yang belum dilakukan *preprocessing*, data yang telah dilakukan

preprocessing menggunakan *corelation-based feature selection*, dan data yang telah dilakukan *preprocessing* menggunakan *information gain*.

Pada proses 10 *folds cross validation* data yang berjumlah 281 akan dibagi menjadi 10 bagian dimana 9 dari 10 bagian data akan digunakan sebagai data latih dan 1 data digunakan sebagai data uji. Proses ini dilakukan berulang-ulang sebanyak 10 kali sehingga setiap bagian data akan dicoba sebagai data uji. Berikut hasil dari proses pengujian data dan evaluasi data yang ditampilkan pada Tabel IV.

TABEL IV
HASIL PENGUJIAN DAN EVALUASI DATA

Evaluation	10 Folds Cross Validation		
	Data Awal	Attribut Selection CFS	Attribute Selection IG
Accuracy	76.19	76.92	76.19
Precission L	0.78	0.80	0.78
Precission M	0.74	0.74	0.74
Precission H	0.83	0.82	0.83
Precission AVG	0.78	0.79	0.78
Recall L	0.77	0.81	0.77
Recall M	0.76	0.78	0.76
Recall H	0.74	0.72	0.74
Recall AVG	0.76	0.77	0.76

C. Pembahasan

Berdasarkan Tabel IV dapat dilihat bahwa hasil pengujian dan evaluasi data yang telah diproses menggunakan *10 folds cross-validation* dan algoritma C4.5 yang dikombinasikan dengan metode seleksi fitur *correlation-based feature selection* menghasilkan nilai akurasi yang lebih tinggi. Hal ini dikarenakan pada proses pengujian data, data yang di *preprocessing* menggunakan *correlation-based* dapat menghasilkan 216 data yang di prediksi benar dan 65 salah yang hasilnya dapat dilihat pada Tabel V.

TABEL V
CONFUSION MATRIX PREPROCESSING CORRELATION-BASED

		Predict Class		
		A	*B*	*C*
Actual Class	***L***	50	12	0
	M	13	103	16
	H	0	24	63

216 data yang diperoleh berasal dari penjumlahan tabel *confusion matrix* berwarna oranye. Sementara itu, 65 data yang salah diperoleh dari penjumlahan nilai tabel selain warna oranye.

Data awal yang tidak di *preprocessing* menghasilkan 214 data yang di prediksi benar dan 67 data salah yang dapat dilihat pada Tabel VI.

TABEL VI
MODEL CONFUSION MATRIX PREPROCESSING INFORMATION GAIN

		Predict Class		
		A	***B***	***C***
Actual Class	***L***	48	13	1
	M	12	101	19
	H	1	21	65

214 data yang diperoleh berasal dari penjumlahan tabel *confusion matrix* berwarna kuning. Sementara itu, 67 data yang salah diperoleh dari penjumlahan nilai tabel selain warna kuning.

Data yang di *preprocessing* menggunakan *information gain* menghasilkan 214 data yang di prediksi benar dan 67 data salah yang dapat dilihat pada Tabel VII.

TABEL VII
MODEL CONFUSION MATRIX DATA AWAL TAMPA PREPROCESSING

		Predict Class		
		A	***B***	***C***
Actual Class	***L***	48	13	1
	M	12	101	19
	H	1	21	65

Dari pengujian data tersebut menunjukan bahwa data yang di *preprocessing* menggunakan *correlation-based* menghasilkan selisih 2 data yang dapat di prediksi benar dari data yang tidak di *preprocessing* dan data yang di *preprocessing* menggunakan *information gain*. Hal ini merupakan salah satu pengaruh yang menjadikan nilai akurasi dari data yang di *preprocessing* menggunakan *correlation-based* menjadi lebih tinggi. Pada nilai presisi

kelas H dari ketiga model data menghasilkan nilai paling tinggi. Pada nilai *Recall*, kelas L merupakan nilai yang paling tinggi dari ketiga model data.

Hasil dari pengujian data antara data yang diproses menggunakan algoritma C4.5 tanpa *preprocessing* dan data yang telah di *preprocessing* menggunakan *information gain* menghasilkan nilai akurasi, presisi, dan *recall* yang sama, serta model *tree* dan *rule* yang dibuat juga sama. Hal ini dikarenakan *preprocessing* menggunakan *information gain* dan algoritma C4.5 sama-sama menghitung nilai *gain* untuk membuat model *tree* yang akan digunakan.

Penggunaan akses memory dalam pemrosesan data awal yang tanpa dilakukan *preprocessing* membutuhkan memory sebesar 380.1MB, sedangkan data dengan *preprocessing* menggunakan *information gain* membutuhkan memory yang sedikit lebih besar yaitu, 413.4MB. Sedangkan untuk data dengan *preprocessing* menggunakan *correlation-based* membutuhkan memory sebesar 399.4MB. Hal ini membuktikan bahwa pemrosesan dari data yang tidak dilakukan *preprocessing* membutuhkan akses memory yang lebih sedikit dibandingkan dengan data yang di *preprocessing* menggunakan *information gain* dan *correlation-based*.

Selanjutnya, berdasarkan pengujian waktu pemrosesan data yang dilakukan menggunakan *stopwatch*, data awal tanpa dilakukan *preprocessing* membutuhkan waktu sebanyak 05.3 detik. Untuk data dengan *preprocessing* menggunakan *information gain* membutuhkan waktu sebanyak 05.4 detik. Selanjutnya untuk data dengan *preprocessing* menggunakan *correlation-based* membutuhkan waktu sebanyak 05.1 detik. Hal ini menunjukan bahwa data yang di *preprocessing* menggunakan *correlation-based* merupakan data dengan waktu pemrosesan paling cepat.

IV. KESIMPULAN

Berdasarkan komparasi dari hasil pengujian 3 model data yang berbeda menunjukan bahwa data dengan *preprocessing* menggunakan *correlation-based* menghasilkan tingkat akurasi yang lebih tinggi dari data yang tidak di *preprocessing* dan data yang di *preprocessing* menggunakan *information gain*.

Pada nilai presisi kelas H merupakan kelas dengan nilai paling tinggi dari ketiga model data. Pada nilai *Recall*, kelas L merupakan nilai yang paling tinggi dari ketiga model data.

Hasil dari pengujian data antara data yang diproses menggunakan algoritma C4.5 tanpa *preprocessing* dan data yang telah di *preprocessing* menggunakan *information gain* menghasilkan nilai akurasi yang sama.Hasil dari pengujian waktu pemrosesan data, data yang di *preprocessing* menggunakan *correlation-based feature selection* sedikit lebih cepat dengan waktu pemrosesan selama 05.1 detik. Hasil untuk penggunaan akses memory dalam pemrosesan data, data yang tidak dilakukan *preprocessing* membutuhkan akses memory yang lebih sedikit dibandingkan data yang di *preprocessing* menggunakan *information gain* dan *correlation-based feature selection*.

REFERENSI

[1] Y. Yamasari, S. M. S. Nugroho, D. F. Suyatno, and M. H. Purnomo, "Meta-Algoritme Adaptive Boosting untuk Meningkatkan Kinerja Metode Klasifikasi pada Prestasi Belajar Mahasiswa," vol. 6, no. 3, 2017.

[2] dkk Sarthika, "Analisis Profil Mahasiswa Politeknik Negeri Batam dengan Teknik Data Mining Asosiasi dan Clustering," vol. 8, no. 1, pp. 16–21, 2016.

[3] B. Novianti, T. Rismawan, and S. Bahri, "Implementasi Data Mining dengan Algoritma C4 . 5 untuk Penjurusan Siswa (Studi Kasus : SMA Negeri 1 Pontianak)," vol. 04, no. 3, 2016.

[4] A. Maulana, M.F., dan Karomi, M., "Information Gain untuk Mengetahui Pengaruh Atribut Terhadap Klasifikasi Persetujuan Kredit," vol. 9, 2015.

[5] T. dan M. Djatna, "Pembandingan Stabilitas Algoritma Seleksi Fitur menggunakan Transformasi Ranking Normal," 2015.

[6] B. N. Sari, "Implementasi Teknik Seleksi Fitur Information Gain pada Algoritma Klasifikasi Machine Learning untuk Prediksi Performa Akademik Siswa," pp. 6–7, 2016.

[7] I. Rahmayuni, "Perbandingan Performansi Algoritma C4.5 dan Cart Dalam Klasifiksi Data Nilai Mahasiswa Prodi Teknik komputer Politeknik Negeri Padang," *J. Teknol. Inf. Pendidik.*, vol. 2, no. 1, pp. 87–94, 2014.

[8] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods," *Int. J. Database Theory Appl.*, vol. 9, no. 8, pp. 119–136, 2016.

[9] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's performance," *2015 IEEE Jordan Conf. Appl. Electr. Eng. Comput. Technol.*, no. November, pp. 1–5, 2015.

[10] E. Juliyanto, S. E. Nugroho, and P. Marwoto, "Perkembangan Pola Pemecahan Masalah Anak The Pattern of Development of School-Age Children Troubleshooting Solving Problems In," vol. 9, pp. 151–162, 2013.

[11] J. Alfin, "Analisis karakteristik siswa pada tingkat sekolah dasar," pp. 190–205, 2015.

[12] M. Doshi, S. K. Chaturvedi, and D. Ph, "Correlation Based Feature Selection (CFS) Technique to Predict Student Perfromance," vol. 6, no. 3, pp. 197–206, 2014.

[13] I. Y. Purbasari and B. Nugroho, "Benchmarking Algoritma Pemilihan Atribut Pada Klasifikasi Data Mining," *Snastia*, pp. 47–54, 2013.

[14] dkk I. Maulida, "Seleksi Fitur Pada Dokumen Abstrak Teks Bahasa Indonesia Menggunakan Metode Information Gain," no. October 2016, 2017.

[15] Sujana, "Aplikasi mining data mahasiswa dengan metode klasifikasi decision tree," vol. 2010, no. Snati, 2010.

[16] A. R. Khadafy, "Penerapan Naive Bayes untuk Mengurangi Data Noise pada Klasifikasi Multi Kelas dengan Decision Tree," vol. 1, no. 2, pp. 136–142, 2015.